政治学方法論 II 課題 2

宋財泫 (SONG Jaehyun) (123J009J)

2015-4-15(水)

1 確率

1. 以下の関係が成り立つことを示しなさい。

$$E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

X をスカラーの確率変数、$E(X)$ を確率変数 X の期待値だとする。

$$\begin{aligned} E[X - E(X)]^2 &= E[X^2 - 2E(X)X + [E(X)]^2] \\ &= E(X^2) - E[2E(X)X] + E[[E(X)]^2] \end{aligned}$$

E(X) は平均値であり、定数 (constant) なので $E(c) = c$ の関係が成立する。したがって、

$$\begin{aligned} E[X - E(X)]^2 &= E(X^2) - E[2E(X)X] + E[[E(X)]^2] \\ &= E(X^2) - \underbrace{2E(X)E(X)}_{2[E(X)]^2} + [E(X)]^2 \\ &= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

2. 迷惑メールフィルタは、迷惑メールに頻繁に登場する単語またはフレーズによってメールの判別を行う。世の中の 75% のメールが迷惑メールだと仮定しよう。さらに、迷惑メールのうち 10% には「当選おめでとうございます」というフレーズが登場し、同じフレーズは迷惑メールでなければ 3% にしか登場しないとする。新着

のメールに「当選おめでとうございます」というフレーズがあるとき、このメールが迷惑メールである確率を求めなさい。

$P(Spam) = 0.75$：迷惑メールの確率
$P(Atari|Spam) = 0.1$：迷惑メールのうち、「当選おめでとうございます」が登場する確率
$P(Atari|\neg Spam) = 0.03$：通常メールのうち、「当選おめでとうございます」が登場する確率

$$
\begin{aligned}
P(Spam|Atari) &= \frac{P(Atari|Spam)P(Spam)}{P(Atari)} \\
&= \frac{P(Atari|Spam)P(Spam)}{P(Atari|Spam)P(Spam) + P(Atari|\neg Spam)P(\neg Spam)} \\
&= \frac{0.1 \times 0.75}{0.1 \times 0.75 + 0.03 \times 0.25} \\
&= \frac{0.075}{0.075 + 0.0075} = \frac{0.075}{0.0825} \\
&= 0.90909...
\end{aligned}
$$

新着のメールに「当選おめでとうございます」というフレーズがあるとき、このメールが迷惑メールである確率は約 90.9% である。

3. 女の双子を妊娠中の女性がいるとする。双子は一卵性または二卵性である。一般的に、双生児のうちの 1/3 が一卵性である。一卵性双生児の性別は必ず同じだが、二卵性の場合には性別が同じとは限らない。一卵性であればどちらの性別も同様に確からしく、二卵性の場合には起こり得るすべての組み合わせが同様に確からしいとする。このとき、この女性のお腹の中にいる双子が一卵性である確率を求めなさい。

$P(I) = 1/3$：一卵性の確率
$P(F) = 2/3$：一卵性の確率
$P(G, G|I) = 1/2$：一卵性のとき、二人の性別が女 (*G*irl) である確率

$$
\begin{aligned}
P(I|G,G) &= \frac{P(G,G|I)P(I)}{P(G,G)} \\
&= \frac{P(G,G|I)P(I)}{P(G,G|I)P(I) + P(G,G|F)P(F)}
\end{aligned}
$$

$P(G,G|F)$ は二卵性のとき、可能な性別の組合せは (G,G)、(B,B)、(G,B)、(B,G) である (B は男 Boy を意味する)。したがって、$P(G,G|F)$ は $1/4$ になる。

$$
\begin{aligned}
P(I|G,G) &= \frac{P(G,G|I)P(I)}{P(G,G|I)P(I) + P(G,G|F)P(F)} \\
&= \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1/4 \times 2/3} \\
&= \frac{1/6}{1/6 + 2/12 = 2/6} \\
&= \frac{1/6}{2/6} = 0.5
\end{aligned} \tag{1}
$$

女の双子を妊娠中の女性のお腹の中にいる双子が一卵性である確率は 50% である。

4. 袋の中に 100 枚のコインがあり、そのうち 99 枚は公平なコイン（確率 1/2 で表が出る)、残りの 1 枚はマジック用コイン（必ず表が出る：両面とも表）である。袋の中から無作為に 1 枚のコインを選び、選んだコインを 7 回投げたところ、7 回とも表が出た。このとき、選んだコインがマジック用のコインである確率を求めなさい。（ただし、コインの両面を直接観察して判断することはできない。）

   $P(M) = 0.01$：マジック・コインが選ばれる確率
   $P(F) = 0.99$：公平なコインが選ばれる確率
   $P(H7|M) = 1$：マジック・コインが選ばれたとき、7 回中 7 回が表である確率

コインを 7 回投げたところ、7 回とも表が出たとき、選んだコインがマジック用のコインである確率、$\frac{P(H7|M)P(M)}{P(H7)}$ は、

$$
P(M|H7) = \frac{P(H7|M)P(M)}{P(H7)}
$$

であるが、$P(H7)$、とりわけ $P(H7|F)$ が問題にある。公平なコイン ($\theta = 0.5$) を選んだと仮定し、7 回投げて 7 回表が出る確率は、

$$
\begin{aligned}
P(H7|F) &= \binom{7}{7}\theta^7(1-\theta)^{(7-7)} \\
&= 1 \times 0.5^7 \times 1 \\
&= 0.0078125
\end{aligned} \tag{2}
$$

である。したがって $\frac{P(H7|M)P(M)}{P(H7)}$ は、

$$
\begin{aligned}
P(M|H7) &= \frac{P(H7|M)P(M)}{P(H7)} \\
&= \frac{P(H7|M)P(M)}{P(H7|M)P(M) + P(H7|F)P(F)} \\
&= \frac{1 \times 0.01}{1 \times 0.01 + 0.0078125 \times 0.99} \\
&= \frac{0.01}{0.01 + 0.007734375} = \frac{0.01}{0.017734375} \\
&= 0.56387...
\end{aligned}
$$

である。したがって、コインを 7 回投げたところ、7 回とも表が出たとき、選んだコインがマジック用のコインである確率は約 56.39% である。

2　誕生日問題のシミュレーション

以下の条件でシミュレーションを行いなさい。最終的な答えだけでなく、自分で設定した条件を含めたシミュレーションの過程も説明すること。(ヒント：理論値の計算には pbirthday() 関数を使う)

1. 10 人の集団に誕生日を無作為に割り当て、少なくとも 1 組（2 人）が同じ誕生日になる確率を、シミュレーションによる相対頻度として求め、理論値と比較しなさい。

pbirthday() 関数を使い、10 人の集団に誕生日を無作為に割り当て、少なくとも 1 組（2 人）が同じ誕生日になる確率理論値を求める。

```
> pbirthday(10, classes=365, 2)
[1] 0.1169482
```

続いて、シミュレーションのための関数 (bday.sim) を作成する。関数が以下のようなアルゴリズムで設計する

1. 引数は people(人数), coincide(同じ誕生日の人の数の条件), trials(試行回数)
2. シミュレーションの ID(id) と coincide 以上の人が同じ誕生日であったか否か (max) を書き込むデータフレーム (result.df) を作成する。
3. 1 から 365 までの数字の中で people の数だけ反復抽出し、そのベクトルをオブジェクト b.day に書き込む。
4. table(b.day) はベクトル b.day の度数分布表であり、max(table(b.day)) は最大度数を表示する。たとえば max(table(b.day)) が 3 だと、いずれかの日で 3 人が同じ誕生日だということを意味する。ただし、max(table(b.day)) だけでは、それが何月何日かは分からない。
5. もし max(table(b.day)) が coincide と同じかあるいは大きかったら result.df に ID と 1 を、そうでなかったら ID と 0 を書き込む。
6. 以上の過程を trials の数だけ反復する。
7. 最終的には result.df の max 列の総和を trials で割った値を返還する。

以上の手順に基づいたシミュレーション関数のソースは以下のとおりである。

```
bday.sim <- function(people = 10, coincide = 2, trials = 10000)
    {
  result.df <- data.frame(id = rep(NA, trials),
                          max = rep(NA, trials))
  n <- people
  for(i in 1:trials){
    b.day <- sample(x = 1:365, size = n, replace = TRUE)
    if(max(table(b.day)) >= coincide){
      result.df[i, ] <- c(i, 1)
    }else{
      result.df[i, ] <- c(i, 0)
    }
  }
  return(sum(result.df[, 2])/trials)
}

> bday.sim(people = 10, coincide = 2, trials = 10000)
[1] 0.1189
```

1 万回のシミュレーションの結果、理論値 (0.1169482) と近い 0.1189 が得られた。

2. 40 人の集団に誕生日を無作為に割り当て、少なくとも 1 組（2 人）が同じ誕生日になる確率を、シミュレーションによる相対頻度として求め、理論値と比較しなさい。

```
> pbirthday(n = 40, classes = 365, coincident = 2)
[1] 0.8912318

> bday.sim(people = 40, coincide = 2, trials = 10000)
[1] 0.8885
```

1 万回のシミュレーションの結果、理論値 (0.8912318) と近い 0.8885 が得られた。

3. 40 人の集団に誕生日を無作為に割り当て、少なくとも 3 人が同じ誕生日になる確率を、シミュレーションによる相対頻度として求め、理論値と比較しなさい。

```
> pbirthday(n = 40, classes = 365, coincident = 3)
[1] 0.07112002

> bday.sim(people = 40, coincide = 3, trials = 10000)
[1] 0.0706
```

1 万回のシミュレーションの結果、理論値 (0.07112002) と近い 0.0706 が得られた。

以上の結果をまとめると、シミュレーションからも理論値とかなり近い値が得られることが確認できる。実際にはややズレがあるが、これは試行回数をより増やすことによってより理論値に近似させることが出来ると考えられる。